Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images

Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, Gregory Rogez

Moulding Humans: Non-parametric 3D Human Shape Estimation from Single Images

Valentin Gabeur[1,2] Jean-Sébastien Franco[1] Xavier Martin[1] Cordelia Schmid[1,2] Grégory Rogez[3,†]
[1] Inria* [2] Google Research [3] NAVER LABS Europe

Abstract

In this paper, we tackle the problem of 3D human shape estimation from single RGB images. While the recent progress in convolutional neural networks has allowed impressive results for 3D human pose estimation, estimating the full 3D shape of a person is still an open issue. Model-based approaches can output precise meshes of naked under-cloth human bodies but fail to estimate details and un-modelled elements such as hair or clothing. On the other hand, non-parametric volumetric approaches can potentially estimate complete shapes but, in practice, they are limited by the resolution of the output grid and cannot produce detailed estimates. In this work, we propose a non-parametric approach that employs a double depth map to represent the 3D shape of a person: a visible depth map and a "hidden" depth map are estimated and combined, to reconstruct the human 3D shape as done with a "mould". This representation through 2D depth maps allows a higher resolution output with a much lower dimension than voxel-based volumetric representations. Additionally, our fully derivable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the 3D output. We train and quantitatively validate our approach on SURREAL and on 3D-HUMANS, a new photorealistic dataset made of semi-synthetic in-house images annotated with 3D ground truth surfaces.

1. Introduction

Recent works have shown the success of deep network architectures for the problem of retrieving 3D features such as kinematic joints [4, 33] or surface characterizations [43] from single images, with extremely encouraging results. Such successes, sometimes achieved with simple, standard network architectures [30], have naturally motivated the applicability of these methodologies for the more challenging problem of end-to-end full 3D human shape retrieval [2, 18]. The ability to retrieve such information from single images or videos is relevant to a broad number of applications, from self-driving cars, where spatial understanding of surrounding obstacles and pedestrians plays a key role, to animation or augmented reality applications such as virtual change rooms that can offer the E-commerce industry a virtual fitting solution for clothing or bodywear.

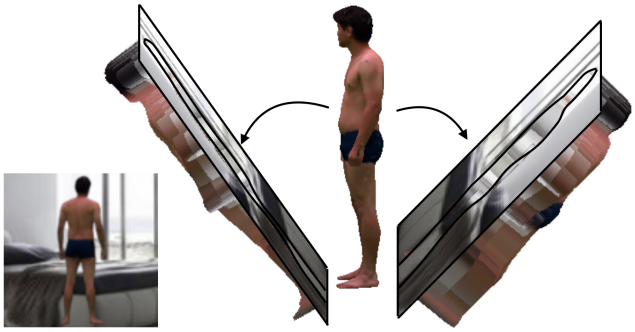


Figure 1. Our non-parametric representation for human 3D shape: given a single image, we estimate the "visible" and the "hidden" depth maps from the camera point of view. The two depth maps can be seen as the two halves of a virtual "mould". We show this representation for one of the images of our new dataset.

Designing a deep architecture that produces full 3D shapes of humans observed in an input image or a sequence of input images raises several key challenges. First, there is a representational issue. While the comfort zone of CNNs is in dealing with regular 2D input and output grids, the gap must be bridged between the 2D nature of inputs and the 3D essence of the desired outputs. One solution is to follow a parametric method and estimate the deformation parameters of a predefined human 3D model [2, 18]. These methods are limited to the level of details covered by the model. In contrast, non parametric approaches can potentially account for shape surface details but are prone to produce physically-impossible body shapes. This is the case of the recent volumetric approach proposed in [38] that encodes the human body as a voxel grid whose dimensionality directly impacts

* Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France. † Most of the work was done while the last author was a research scientist at Inria.

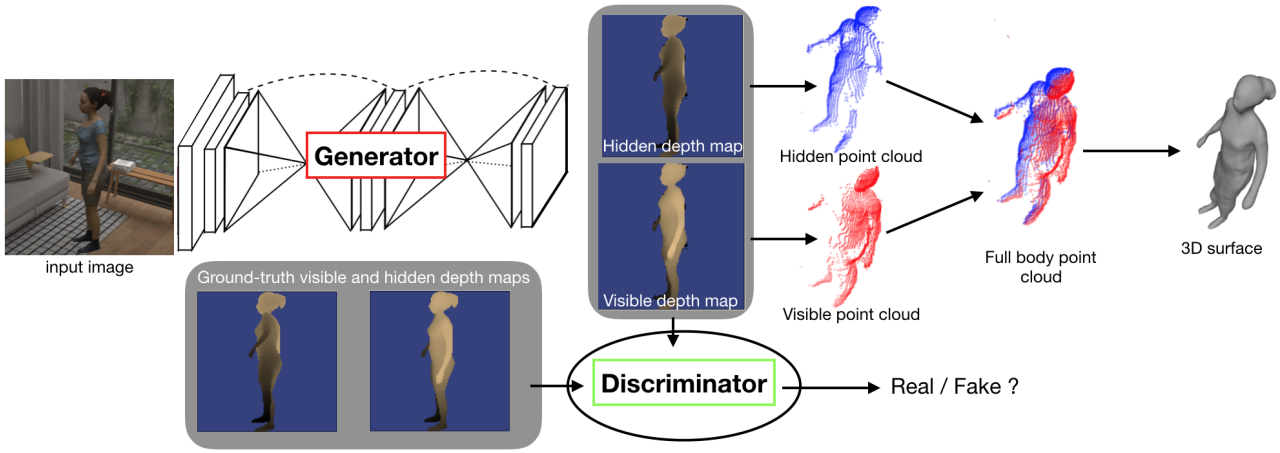


Figure 2. Overview. Given a single image, we estimate the "visible" and the "hidden" depth maps. The 3D point clouds of these 2 depth maps are combined to form a full-body 3D point cloud, as if lining up the 2 halves of a "mould". The 3D shape is then reconstructed using Poisson reconstruction [19]. An adversarial training with a discriminator is employed to increase the humanness of the estimation.

the precision of the estimation. This highlights a second challenge: the dimensionality of the problem is considerably higher than what existing networks have been shown to handle, because the parametrisation sought is no longer restricted to a subset of the variability, e.g. kinematic pose of humans or body shape parameters, but to an intrinsically finer description of the body. Finally, the training data for this problem, yet to be produced, requires a particularly demanding definition and acquisition effort. The large data variability of 3D problems has motivated some initial efforts to produce fully synthetic training sets [39], where such variability can be partially scripted. Yet recent successful methods underscore the necessity for realistic data, for both the general applicability of the estimation, and to keep the underlying network architecture simple, as devoid as possible of any domain specific adaptations.

In order to overcome these difficulties, we propose a non-parametric approach that employs a double depth map representation to encode the 3D shape of a person: a "visible" depth map capturing the observable human shape and a "hidden" depth map representing the occluded surface are estimated and combined to reconstruct the full human 3D shape. In this encoding of the 3D surface, the two depth maps can be seen as the two halves of a virtual "mould", see Figure 1. This representation allows a higher resolution output, potentially the same as the image input, with a much lower dimension than voxel-based volumetric representations, i.e. $O(N^2)$ vs $O(N^3)$. We designed an encoder-decoder architecture that takes as input a single image and simultaneously produces an estimate for both depth maps. These depth maps are then combined to obtain a point cloud of the full 3D surface which can be readily reconstructed using Poisson reconstruction [19]. Importantly, our fully differentiable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the 3D output, especially in the case of strong occlusions. See Figure 2. To train and quantitatively evaluate our network in near real-world conditions, we captured a large-scale dataset of textured 3D meshes that we augment with realistic backgrounds. To account for the large variability in human appearance, we took special care in capturing data with enough variability in movements, clothing and activities. Compared to parametric methods, our method can estimate detailed 3D human shapes including hair, clothing and manipulated objects.

After reviewing the related work in Section 2, we present our two-fold contribution: our new non-parametric 3D surface estimation method is explained in Section 3 while our large-scale dataset of real humans with ground-truth 3D data is detailed in Section 4. Experiments are presented in Section 5 and conclusions drawn in Section 6.

2. Related Work

3D object from single images. Various representations have been adopted for 3D object shape estimation. Voxel-based representations [5] consist in representing the 3D shape as an occupancy map defined on a fixed resolution voxel grid. Octree methods [36] improve the computability of volumetric results by reducing the memory requirements. Point-clouds are another widely employed representation for 3D shapes. In [7], Fan et al. estimate sets of 1024 points from single images. Jiang et al. [15] build on this idea and incorporate a geometric adversarial loss (GAL) to improve the realism of the estimations. AtlasNet [10] directly estimates a collection of parametric surface elements to represent a 3D shape. Our representation combines two complementary depth maps aligned with the image, similar in spirit to the two halves of a "mould", and shares the resolution of the input image, capturing finer details while keeping output dimensionality reasonable.

Similarly to the work of Tatarchenko et al. [35] on reconstructing vehicle images from different viewpoints, we combine the estimation of several depth maps to obtain a 3D shape. For human shape estimation, however, we work on a deformable object. Also, we focus on the visible and hidden depth maps rather than any other because of their direct correspondence with the input image. Our two depth maps being aligned with the image, details as well as contextual image information are directly exploited by the skip connections to estimate the depth values. Multi-views [35] do not necessarily have pixel-to-pixel correspondences with the image making depth prediction less straightforward.

3D human body shape from images. Most existing methods for body shape estimation from single images rely on a parametric model of the human body whose pose and shape parameters are optimized to match image evidence [2, 11, 20, 29]. This optimization process is usually initialised with an estimate of the human pose supplied by the user [11] or automatically obtained through a detector [2, 20, 29] or inertial sensors [42]. Instead of optimizing mesh and skeleton parameters, recent approaches proposed to train neural networks that directly predict 3D shape and skeleton configurations given a monocular RGB video [37], multiple silhouettes [6] or a single image [18, 25, 28]. Recently, BodyNet [38] was proposed to infer the volumetric body shape through the generation of likelihoods on the 3D occupancy grid of a person from a single image.

A large body of work exists to extract human representations from multiple input views or sensors, of which some recently use deep learning to extract 3D human representations [8, 13, 21]. While they intrinsically aren't designed to deal with monocular input as proposed, multi-view methods usually yield more complete and higher precision results as soon as several viewpoints are available, a useful feature we leverage for creating the 3D HUMANS dataset.

More similar to ours are the methods that estimate projections of the human body: in [39], an encoder-decoder architecture predicts a quantized depth map of the human body while in DensePose [12] a mapping is established between the image and the 3D surface. Our method also makes predictions aligned with the input image but the combination of two complementary "visible" and "hidden" depth maps leads to the reconstruction of a full 3D volume. In [24], the authors complete the 3D point cloud built from the front facing depth map of a person in a canonical pose by estimating a second depth map of the opposite viewpoint. We instead predict both depth maps simultaneously from a single RGB image and consider a much wider range of body poses and camera views. All these methods rely on a parametric 3D model [2, 18, 20] or on training data annotated [12] or synthesised [39] using such a model. These models of humans built from thousands of scans of naked people such as the SMPL model [23] lack realism in terms of appearance. We instead propose to tackle real-world situations, modeling and estimating the detailed 3D body shape including clothes, hair and manipulated objects.

3D human datasets. Current approaches for human 3D pose estimation are built on deep architectures trained and evaluated on large datasets acquired in controlled environments with Motion Capture systems [1, 14, 34]. However, while the typology of human poses on these datasets captures the space of human motions very well, the visual appearance of the corresponding images is not representative of the scenarios one may find in unconstrained real-world images. There has been a recent effort to generate in-the-wild data with ground truth pose annotation [26, 32]. All these datasets provide accurate 3D annotation for a small set of body keypoints and ignore 3D surface with the exception of [20] and [41] who annotate the SMPL parameters in real-world images manually or using IMU. Although the resulting dataset can be employed to evaluate under-cloth 3D body shapes, its annotations are not detailed enough, and importantly, its size is not sufficient to train deep networks.

To compensate for the lack of large scale training data required to train CNNs, recent work has proposed to generate synthetic images of humans with associated ground truth 3D data [4, 31, 39]. In particular, the Surreal dataset [39], produced by animating and rendering the SMPL model [23] on real background images, has proven to be useful to train CNN architectures for body parts parsing and 2.5D depth prediction [39], 3D pose estimation [31, 33], or 3D shape inference [38]. However, because it is based on the SMPL model, this dataset is not realistic in terms of clothing, hair or interactions with objects and cannot be used to train architectures that target the estimation of a detailed 3D human shape. We propose to bridge this gap by leveraging multi-camera shape data capture techniques [3, 40], introducing the first large scale dataset of images showing humans in realistic scenes, i.e. wearing real clothes and manipulating real objects, dedicated to training with full 3D mesh and pose ground-truth data. Most similar to ours are the CMU Panoptic dataset [17] that focus on social interactions and the data of [45] that contains dense unstructured geometric motion data for several dressed subjects.

3. Methodology

In this section, we present our new non-parametric 3D human shape representation and detail the architecture that we designed to estimate such 3D shape from a single image.

3.1. "Mould" representation

We propose to encode the 3D shape of a person through a double 2.5D depth map representation: a "visible" depth map that depicts the elements of the surface that are directly observable in the image, and a "hidden" depth map that characterises the occluded 3D surface. These two depth

maps can be estimated and combined to reconstruct the complete human 3D shape as done when lining up the two halves of a "mould". See example in Figure 2.

Given a 3D mesh, obtained by animating a 3D human model or by reconstructing a real person from multiple views, and given a camera hypothesis, i.e. location and parameters, we define our two 2D depth maps z_{vis} and z_{hid} by ray-tracing. Specifically, we cast a ray from the camera origin, in the direction of each image pixel location (u, v) and find the closest intersecting point on the mesh surface:

$$z_{vis}[u, v] = \min_{k \in \text{Ray}(u,v)} ||\mathbf{p}_k||_2 \quad (1)$$

for the visible map, and the furthest one for the hidden map:

$$z_{hid}[u, v] = \max_{k \in \text{Ray}(u,v)} ||\mathbf{p}_k||_2, \quad (2)$$

where 3D points $\{\mathbf{p}_i\} = \{(p_{x,i}, p_{y,i}, p_{z,i})\}$ are expressed in camera coordinate system and the L2-norm $||.||_2$ is the distance to the camera center. $\text{Ray}(u, v)$ denotes the set of points $\mathbf{p}_i$ on the ray passing through pixel (u, v) obtained by hidden surface removal and visible surface determination.

To be independent from the distance of the person to the camera, we center the depth values on the center of mass of the mesh, i.e. z_{orig} : $z_{vis}[u, v]' = z_{vis}[u, v] - z_{orig}$ $\forall u, v$, and similarly for $z_{hid}[u, v]$. Since they are defined with respect to the same origin, the 2 depth maps $z_{vis}[u, v]$ and $z_{hid}[u, v]$ can be readily combined in 3D space by merging their respective 3D point clouds into a global one:

$$\{\mathbf{p}_i\} = \{\mathbf{p}_i\}_{vis} \bigcup \{\mathbf{p}_i\}_{hid} \quad (3)$$

An example of such a point cloud is depicted in Figure 2, where points corresponding to $z_{vis}[u, v]$ and $z_{hid}[u, v]$ are respectively colored in red and blue. In practice, to keep the depth values within a reasonable range and estimate them more accurately, we place a flat background a distance L behind the subject to define all pixels values in the depths maps in the range $[-z_{orig} \dots L]$. Points $\mathbf{p}_i$ of the point clouds are then selected as belonging to the human surface if $p_{z,i} \leq L - \epsilon$.

As in volumetric representation through voxel grid, our method also encodes 3D surfaces and point clouds of diverse sizes into a fixed size representation, making a 3D surface easier to consider as a deep network target. However, in our case, we can work at the image resolution with a much lower output dimensionality $O(N^2)$ than voxel-based volumetric representations $O(N^3)$, N being the size of the bounding box framing the human in the input image.

We numerically validated the benefit of our representation compared to a voxel grid approach by encoding a random set of 100 meshes (picked from our 3D HUMANS dataset presented in Section 4) at different resolutions and computing the 3D reconstruction error (average Chamfer distance) between ground-truth vertices and the resulting point clouds. This comparison is shown in Figure 3. The error obtained with our mould-representation decreases and converges to a minimum value that corresponds to surface details that cannot be correctly encoded even with high resolution depth maps, i.e. when some rays intersect more than twice with the human surface for particular poses. In practice, we show in Section 5 that this can be solved by employing a Poisson reconstruction to obtain a smooth 3D surface, including those areas. We can extrapolate from Figure 3 that voxel grids can reach perfect results with an infinity of voxels, but for manageable sizes, our representation allows to capture more details.

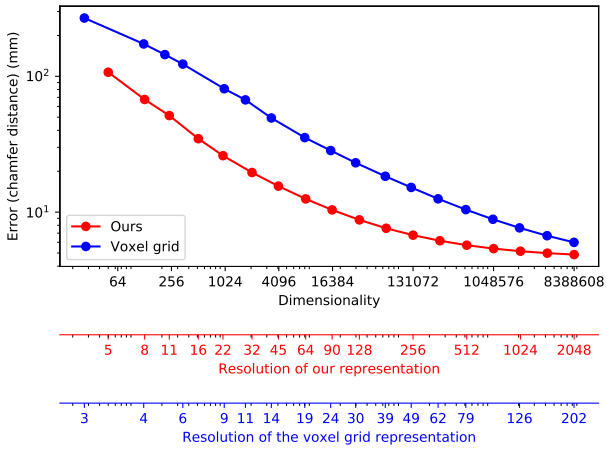


Figure 3. Reconstruction error for voxel grid and our "mould" when augmenting the dimensionality D of the representation, D=N^3 for voxels grid and D=$2N^2$ for ours.

3.2. Architecture

We formulate the 3D shape estimation problem as a pixel-wise depth prediction task for both visible and hidden surfaces. Our framework builds on the stacked hourglass network proposed by Newell et al. [27] that consists of a sequence of modules shaped like an hourglass, each taking as input the prediction from the previous module. Each of these modules has a set of convolutional and pooling layers that process features down to a low resolution and then upsample them until reaching the final output resolution. This process, combined with intermediate supervision through skip connections, implicitly captures the entire context of the image. Originally introduced for the task of 2D pose estimation and employed later for part segmentation and depth prediction [39], this network is an appropriate choice as it predicts a dense pixel-wise output while capturing spatial relationships associated with the entire human body.

We designed a 2-stack hourglass architecture that takes as input an RGB image I cropped around the human and outputs the 2 depths maps z_{vis} and z_{hid} aligned with I. We

use a $\mathcal{L}_{L1}$ loss function defined on all pixels of both depth maps. The loss function to be minimized is thus the average distance between the ground truth z_p and the estimation $\widehat{z_p}$:

$$\mathcal{L}_{L1} = \frac{1}{P}\sum_{p=1}^{P}|z_p - \widehat{z_p}|, \tag{4}$$

with P being the number of pixels in the batch and $\widehat{z_p}$ the network output for pixel p, including pixels in both $z_{vis}[u,v]$ and $z_{hid}[u,v]$ maps.

We also experimented with an $\mathcal{L}_{L2}$ loss but found that it overly penalizes outliers, i.e. pixels incorrectly assigned to background and vice versa, and therefore focuses only on that task. By using the $\mathcal{L}_{L1}$ norm, we force the network to not only segment the image correctly, i.e discriminate the subject from the background, but also provide an accurate estimation of the depth at each pixel.

3.3. Adversarial training

As observed with other non-parametric methods [38] but also with approaches relying on a model [18], our network can sometimes produce implausible shapes that do not look human, especially when a limb is entirely occluded by other parts of the body. To improve the accuracy and the "humanness" of our prediction, we follow an adversarial training procedure inspired by the Generative Adversarial Networks (GAN) [9]. Our fully derivable depth-based model allows us to efficiently incorporate a discriminator in an adversarial fashion, i.e., the goal for the discriminator will be to correctly identify ground truth depth maps from generated ones. On the other hand, the generator objective will be two-fold: fitting the training set distribution through the minimization of the $\mathcal{L}_{L1}$ loss (Equation 4) and tricking the discriminator into classifying the generated depth maps as ground truth depth maps through the minimization of the $\mathcal{L}_{GAN}$ loss:

$$\mathcal{L}_{GAN}(G,D) = E_{I,z}[\log D(I,z)] + E_I[\log(1 - D(I,G(I))]. \tag{5}$$

Our discriminator D will be trained to maximize the $\mathcal{L}_{GAN}$ loss by estimating 1 when provided with ground-truth depth maps z and estimating 0 when provided with generated depth maps $G(I)$. In order to weigh the contribution of each loss, we will use a factor λ, our full objective being modeled as a minimax game:

$$(G^*, D^*) = \arg\min_G \max_D (\mathcal{L}_{GAN}(G,D) + \lambda \mathcal{L}_{L1}(G)). \tag{6}$$

The $\mathcal{L}_{L1}$ loss will be used to learn the training set distribution by retrieving the low-frequency coefficients while the $\mathcal{L}_{GAN}$ loss will entice the generator into predicting realistic and precise depth maps. It is important to note that the discriminator is only used to guide the generator during the learning. The discriminator is not used at test time.

The architecture employed as our discriminator is a 4 stack CNN. Each stack is composed of a convolutional layer (kernel size 3, stride 1), a group normalization layer (32 groups), a ReLu activation function and a MaxPool 2x2 operation. There are 64 channels for the first convolution and the number of channels is multiplied by 2 at each stack until reaching 512 for the 4th and last stack convolution. We then connect our 8x8x512 ultimate feature map with 2 fully-connected layers of size 1024 and 512 neurons and then our final output neuron on which we apply a binary cross entropy loss. We jointly trained our generator and discriminator on 50,000 images for 40 epochs. Training is performed on batches of size 8 with the Adam optimizer. Given our small training batch size, we found the use of group norm [44] to be a great alternative to batch norm that was producing training instabilities. The learning rate is kept constant at 1e-4 during the first 20 epochs and is then decreased linearly to zero during the following 20 epochs. In practice, since our $\mathcal{L}_{L1}$ loss is much smaller than the $\mathcal{L}_{GAN}$ loss, we multiply the $\mathcal{L}_{L1}$ loss by a λ factor equal to 1e4. With this adversarial training, we observed that the results are sharper and more realistic. In cases of deformed or missing limb, e.g. the legs in Figure 7 right, the use of a discriminator forces the generator to produce a better prediction.

4. Dataset generation

We introduce 3D HUMANS (HUman Motion, Activities aNd Shape), a realistic large-scale dataset of humans in action with ground-truth 3D data (shape and pose). It consists of semi-synthetic videos with 3D pose and 3D body shape annotations, as well as 3D detailed surface including cloths and manipulated objects. First, we captured 3D meshes of humans in real-life situations using a multi-camera platform. We then rendered these models on real-world background scenes. See examples in Figure 4a.

Capture. We employed a state of the art 3D capture equipment with 68 color cameras to produce highly detailed shape and appearance information with 3D textured meshes. The meshes are reconstructed frame by frame independently. They are not temporally aligned and do not share any common topology. We divided the capture into 2 different subsets: in the first one, 13 subjects (6 male and 7 female) were captured with 4 different types of garments (bathing suit/tight clothing, short/skirt/dress, wide cloths and jacket) while performing basic movements e.g., walk, run, bend, squat, knees-up, spinning. In the second subset, 6 subjects, 4 male and 2 female, were captured while performing 4 different activities (talking on the phone, taking pictures, cleaning a window, mopping the floor) in 2 different ways: standing/sitting for talking on the phone, standing/kneeling for taking picture, etc. More than 150k meshes were reconstructed. The dataset was collected at Inria from consenting and informed participants.

(a) (b)

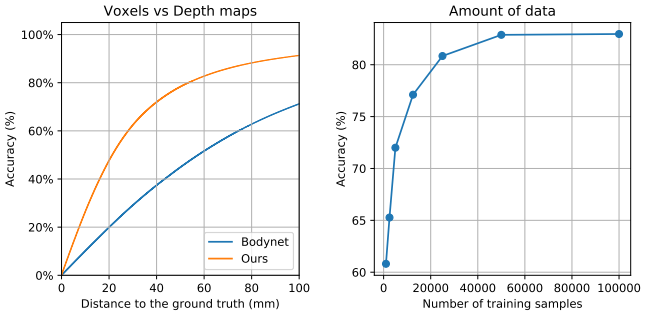
Figure 4. Data generation. (a) We captured 3D meshes of humans, wearing real clothes, moving and manipulating objects using a multi-camera platform. We then rendered these models on real-world background scenes and computed ground-truth visible and hidden depths maps. (b) We also generated a test set by rendering our meshes on realistic 3D environments.

Rendering. We rendered all our videos at a 320 x 240 resolution using a camera of sensor size 32mm and focal length 60mm. Our videos are 100 frames in length and start with the subject at the center of the frame. For the first frame of the sequence, the subject is positioned at a distance of 8 meters of the camera, with a standard deviation of 1 meter. We used the images of the LSUN dataset [46] for background.

Annotations. We augment our dataset with ground-truth SMPL pose and body parameters. To do so, we use the Human3.6M [14] environment as a "virtual MoCap room": we render the 3D meshes for which we want to estimate the 3D pose within that environment, generate 4 views using camera parameters and background images from the dataset and estimate the 2D/3D poses by running LCR-Net++, an off-the-shelf 3D pose detector particularly efficient on Human3.6M. An optimum 3D pose is then computed using multi-view 3D reconstruction and used as initialization to fit the SMPL model, estimating pose and shape parameters that better match each mesh. The SMPL model is fitted to the point clouds both for naked and dressed bodies. Keeping the body parameters fixed (obtained from fits in minimal clothing) resulted in a lower performance of the baseline when evaluated against ground truth dressed bodies.

5. Experiments

We analyse quantitatively and compare our approach to the state-of-the-art on two datasets. First, the SURREAL dataset [39], a synthetic dataset obtained by animating textured human models using MoCap data and rendering them on real background images, and our 3D HUMANS dataset introduced in this paper. While SURREAL covers a wider range of movements since it has been rendered using thousands of sequences from [1], our data better covers shape details such as hair and clothing. In the following experiments, both training and test images are tightly cropped around the person using subjects segmentation. The smallest dimension of the image is extended to obtain a square image that is then resized to 256x256 pixels to serve as input for our network. Performance is computed on both 128x128 output depth maps as the distance between each ground truth foreground pixel and its corresponding pixel in the predicted depth map. Background depth L is set at 1.5m.

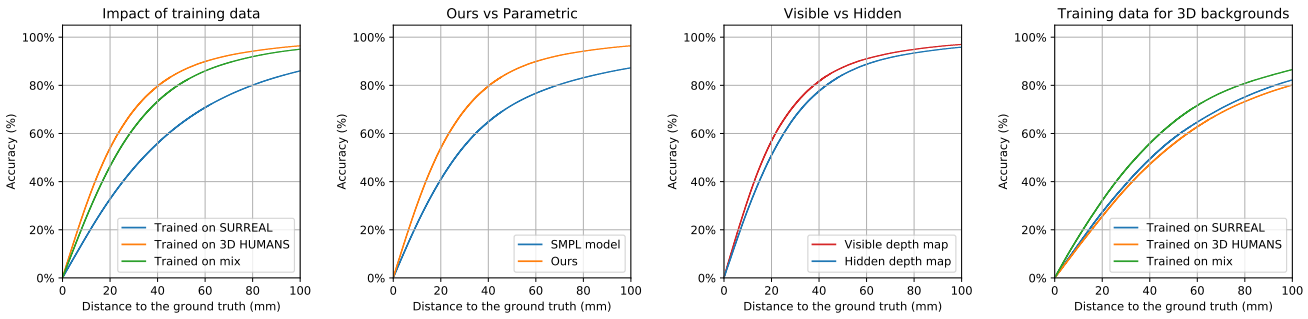


(a) (b)

Figure 5. Comparaison with state-of-the-art on the SURREAL dataset: (a) we first compare against the BodyNet [38] baseline. (b) We analyse the impact of varying the size of the training set on performance on our new 3D HUMANS dataset.

5.1. SURREAL

Recent methods [38, 39] evaluate their performance on this dataset. First, we evaluated the performance of our architecture when estimating quantized depth values (19+1 for background) through classification as in [39] and our proposed regression method: with a maximum distance to groundtruth of 30mm, the quantity of pixels with a correct depth estimation increase by 5% when using regression instead of classification. Then, we compare in Figure 5a our



(a) (b) (c) (d)

Figure 6. Evaluation on our 3D HUMANS dataset: we first analyze the influence of the training data on performance (a). Then, we compare against the SMPL baseline (b). We compare the performance on visible and hidden depth map separately (c). Finally, we analyse the training data on a dataset rendered in realistic backgrounds and observe that SURREAL data is important for generalisation (d).

performance against the recent BodyNet voxel grid-based architecture from [38] who also reported numerical performance on SURREAL. Although good 3D performances are reported in the paper, we can see that when evaluating in the image domain, i.e., comparing depth maps, the performance of BodyNet drops. Our method makes 3D estimations aligned with the image and better recover details, outperforming BodyNet quite substantially.

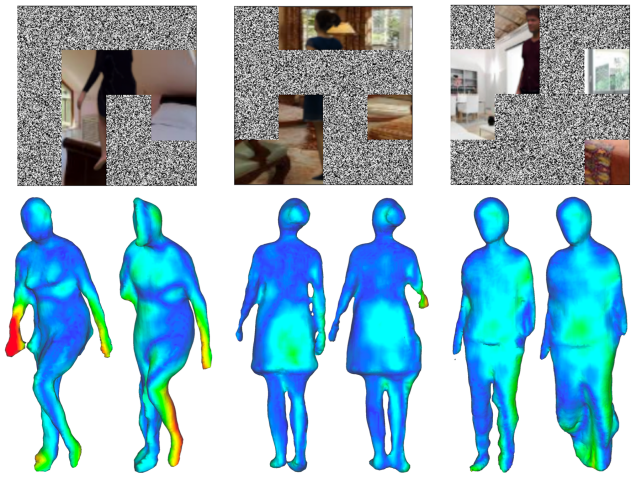


Figure 7. Performance on 3D-HUMANS dataset in presence of severe occlusions on three frames: (top) input images, (left) with GAN, (right) without GAN. Errors above 15cm are shown in red. The GAN helps increase the "humanness" of the predictions.

5.2. 3D HUMANS

We consider 14 subjects (8 male, 6 female) for training and the remaining 5 subjects (2 male, 3 female) for test. An interesting aspect of synthetic datasets is that they offer an almost unlimited amount of training data. In our case, the data generation relies on a capture process with a non-negligible acquisition effort. It is therefore interesting to analyze how adding more training data impacts the performance. Our results in Figure 5b show that training our architecture on 50,000 images is sufficient and that using more training images does not improve much the performance. The appearance of our images being quite different from SURREAL data, we first compare the performance of our method when considering different training strategies: training on SURREAL, training on 3D HUMANS, or training on a mix of both datasets. In Figure 6a, we can see that the best performances are obtained when SURREAL images are not used. The appearance of the images is too different and our architecture cannot recover details such as clothes or hair when trained on data obtained by rendering the SMPL model. This is verified by the result depicted in Figure 6b: we outperform, by a large margin, a baseline obtained by fitting the SMPL model on the ground truth meshes, effectively acting as an upper bound for all methods estimating SMPL meshes [2, 18]. It shows the inefficiency of these methods to estimate clothed body shape since clothes are not included in the SMPL model.

Finally, we analyse in Figure 6c how much our performance varies between front and back depth maps. As expected, we better estimate the visible depth map, but our hidden depth maps are usually acceptable. See examples in Figure 7 and Figure 8. The quality of the 3D reconstructions is remarkable given the low dimensionality of the input. Main failures occur when a limb is completely occluded. In such cases, the network can create non-human shapes. We proposed to tackle this issue by considering an adversarial training that we analyse in the next section. We note a higher performance on 3D HUMANS than on SURREAL. We attribute that to several factors including the higher pose variability in SURREAL (some subjects are in horizontal position) and the absence of lighting in 3D HUMANS. We also analyzed the results on different subsets of the evaluation set @50mm and obtained with/without clothing: 83.30% and 85.43% respectively and with/without object: 79.11% and 84.55% respectively, confirming the nuisance introduced by these elements.

Figure 8. Generalisation to previously unobserved data. We apply our pipeline to images with 3D realistically rendered backgrounds (left), and with 3 real-world images from the LSP dataset (right). These poses, in particular the baseball player, have not been seen at training time but our model still generalizes well.

5.3. GAN

Severe occlusions (self- or by other elements of the scene) are a limitation of our model that we address with adversarial training. We carried out a dedicated experiment where we artificially generated such occlusions in train/test images to quantify improvements. We obtain a 7% chamfer distance error drop with adversarial training and a clear qualitative improvement which we illustrate in Figure 7. We highlight the differences by showing an error heat-map over a Poisson reconstruction of the point cloud for better visualization. The quantitative gain is limited due to the network sometimes hallucinating plausible limbs far from groundtruth (red hand in the left Figure 7), resulting in higher error than a network without GAN that does not estimate any limb at all. This is because the metric does not evaluate the overall plausibility of the produced estimation.

5.4. Generalisation

In order to quantitatively measure its generalisation capability, we have evaluated our network on an additional dataset: instead of static background images, we have rendered the meshes in realistic 3D environments obtained on the internet (examples in Figure 4b). The results (Figure 6d) show that a mix training on both SURREAL and 3D HUMANS is ideal for generalisation. We suspect that jointly rendering the subject and the 3D background at the same time creates a more realistic image where the subject is more complicated to segment, hence the need for more variability in the training data. We also generated qualitative results for LSP images [16], depicted in Figure 8, and for the DeepFashion dataset [22], shown in Figure 9 where we compare our approach with HMR [18] and BodyNet [38]. We can observe that our approach captures more details, including hair, shirt and the belly of the pregnant woman (up), hair, skirt and body pose (middle) and dress (bottom).

Figure 9. Comparison between HMR [18] (left), Bodynet [38] (middle) and our method (right). Unlike [18, 38], we do not train on in-the-wild images but estimate 3D shapes of clothed subjects.

6. Conclusion

We have proposed a new non-parametric approach to encode the 3D shape of a person through a double 2.5D depth map representation: a "visible" depth map depicts the elements of the surface that are directly observable in the image while a "hidden" depth map characterises the occluded 3D surface. We have designed an architecture that takes as input a single image and simultaneously produces an estimate for both depth maps resulting, once combined, in a point cloud of the full 3D surface. Our method can recover detailed surfaces while keeping the output to a reasonable size. This makes the learning stage more efficient. Our architecture can also efficiently incorporate a discriminator in an adversarial fashion to improve the accuracy and "humanness" of the output. To train and evaluate our network, we have captured a large-scale dataset of textured 3D meshes that we rendered on real background images. This dataset will be extended and released to spur further research.

Acknowledgements. We thank Pau De Jorge and Jinlong Yang for their help in capturing the data employed in this paper. The dataset was acquired using the Kinovis[2] platform. This work was supported in part by ERC advanced grant Allegro.

[2]https://kinovis.inria.fr

References

[1] CMU motion capture dataset. http://mocap.cs.cmu.edu. the database was created with funding from nsf eia-0196217. Technical report. 3, 6

[2] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, 2016. 1, 3, 7

[3] D. Casas, M. Volino, J. Collomosse, and A. Hilton. 4D Video Textures for Interactive Character Appearance. *Computer Graphics Forum*, 2014. 3

[4] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3D pose estimation. In *3DV*, 2016. 1, 3

[5] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3D-R2N2: A unified approach for single and multi-view 3D object reconstruction. In *ECCV*, 2016. 2

[6] E. Dibra, H. Jain, A. C. Öztireli, R. Ziegler, and M. H. Gross. Hs-nets: Estimating human body shape from silhouettes with convolutional neural networks. In *3DV*, 2016. 3

[7] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *CVPR*, 2017. 2

[8] A. Gilbert, M. Volino, J. Collomosse, and A. Hilton. Volumetric performance capture from minimal camera viewpoints. In *ECCV*, 2018. 3

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*. 2014. 5

[10] T. Groueix, M. Fisher, V. G. Kim, B. Russell, and M. Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2

[11] P. Guan, A. Weiss, A. O. Balan, and M. J. Black. Estimating human shape and pose from a single image. In *ICCV*, 2009. 3

[12] R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. *arXiv*, 2018. 3

[13] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, C. Ma, L. Luo, and H. Li. Deep volumetric video from very sparse multi-view performance capture. In *ECCV*, 2018. 3

[14] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. PAMI*, 2014. 3, 6

[15] L. Jiang, S. Shi, X. Qi, and J. Jia. GAL: Geometric adversarial loss for single-view 3D-object reconstruction. In *ECCV*, 2018. 2

[16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 8

[17] H. Joo, H. Liu, L. Tan, L. Gui, B. Nabbe, I. Matthews, T. Kanade, S. Nobuhara, and Y. Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 3

[18] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 3, 5, 7, 8

[19] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29:1–29:13, July 2013. 2

[20] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *CVPR*, 2017. 3

[21] V. Leroy, J.-S. Franco, and E. Boyer. Shape Reconstruction Using Volume Sweeping and Learned Photoconsistency. In *ECCV*, 2018. 3

[22] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016. 8

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics*, 2015. 3

[24] N. Lunscher and J. Zelek. Deep learning whole body point cloud scans from a single depth map. In *CVPRW*, 2018. 3

[25] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 3

[26] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, and C. Theobalt. Single-shot multi-person 3D pose estimation from monocular rgb. In *3DV*, sep 2018. 3

[27] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 4, 10

[28] M. Omran, C. Lassner, G. Pons-Moll, P. V. Gehler, and B. Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 3

[29] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. *General Automatic Human Shape and Motion Capture Using Volumetric Contour Cues*. 2016. 3

[30] G. Rogez and C. Schmid. Mocap-guided data augmentation for 3D pose estimation in the wild. In *NIPS*, 2016. 1

[31] G. Rogez and C. Schmid. Image-based synthesis for deep 3D human pose estimation. *IJCV*, 126(9):993–1008, 2018. 3

[32] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *CVPR*, 2017. 3

[33] G. Rogez, P. Weinzaepfel, and C. Schmid. LCR-Net++: Multi-person 2D and 3D Pose Detection in Natural Images. *IEEE Trans. PAMI*, 2019. 1, 3

[34] L. Sigal, A. O. Balan, and M. J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2):4–27, 2010. 3

[35] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Multi-view 3D models from single images with a convolutional network. In *ECCV*, 2016. 3

[36] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3D outputs. In *ICCV*, 2017. 2

[37] H. Tung, H. Tung, E. Yumer, and K. Fragkiadaki. Self-supervised learning of motion capture. In *NIPS*, 2017. 3

[38] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. Bodynet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 1, 3, 5, 6, 7, 8

[39] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid. Learning from Synthetic Humans. In *CVPR*, 2017. 2, 3, 4, 6

[40] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 27(3):97:1–97:9, Aug. 2008. 3

[41] T. von Marcard, R. Henschel, M. Black, B. Rosenhahn, and G. Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 3

[42] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *IEEE Trans. PAMI*, 38(8):1533–1547, 2016. 3

[43] L. Wei, Q. Huang, D. Ceylan, E. Vouga, and H. Li. Dense human body correspondences using convolutional networks. In *CVPR*, 2016. 1

[44] Y. Wu and K. He. Group normalization. In *ECCV*, 2018. 5, 10

[45] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *ECCV*, 2016. 3

[46] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 6

Annex

7. Architecture details

7.1. Generator

The main difference between our generator architecture and the stacked hourglass by Newell et al. [27] is the output dimension. Newell et al. estimate a 64x64 resolution heatmap for each body joint. In our case, we estimate 2 depth maps and aim at a higher 128x128 resolution. Our hourglass output dimension is 128x128x2. Because of this difference in output resolution, we apply the following modifications to the stacked hourglass [27] architecture: We do not use a maxpooling operation after layer1, we increase the depth of the hourglasses from 4 to 5 skipped connections, we project the hourglass result on 2 channels (one for each depth map). Also, we use 2 stacked hourglasses and we replace batch normalization by group normalization [44] that performs better on small training batches. See architecture details in Table 1.

Layer	**Layer type**	**Output shape**
Input	Input	256x256x3
Conv1	Conv 7x7 stride=2, GroupNorm, Relu	128x128x64
Layer1	Residual module expanded	128x128x128
Layer2	Residual module expanded	128x128x256
Layer3	Residual module	128x128x256
Hg1	Hourglass, skipped connections = 5	128x128x2
Hg2	Hourglass, skipped connections = 5	128x128x2

Table 1. Generator architecture.

7.2. Discriminator

For our discriminator, we employed a 4 stacks CNN. It takes as input a set of 2 depth maps at resolution 128x128 and outputs a scalar: close to 1.0 if it believes they are sampled from the ground truth depth maps and close to 0 if it believes they have been generated by the generator. See Table 2 for details.

Layer	**Layer type**	**Output shape**
Input	Input	128x128x2
Conv1	Conv 3x3 stride=1, GroupNorm, Relu	128x128x64
MP1	MaxPool 2x2	64x64x64
Conv2	Conv 3x3 stride=1, GroupNorm, Relu	64x64x128
MP2	MaxPool 2x2	32x32x128
Conv3	Conv 3x3 stride=1, GroupNorm, Relu	32x32x256
MP3	MaxPool 2x2	16x16x256
Conv4	Conv 3x3 stride=1, GroupNorm, Relu	16x16x512
MP4	MaxPool 2x2	8x8x512
FC1	Fully connected layer	1024
FC2	Fully connected layer	512
FC3	Fully connected layer	1

Table 2. Discriminator architecture.